非有界利得をもつセミ・マルコフ決定過程について(学習と制御とその周辺)

AUTHOR(S):

涌田, 和芳

CITATION:

涌田, 和芳. 非有界利得をもつセミ・マルコフ決定過程について(学習と制御とその周辺). 数理解析研究所講究録 1985, 557: 65-78

ISSUE DATE:

1985-04

URL:

http://hdl.handle.net/2433/98985

数理解析研究所講究録
557巻 1985年 65-78

非有界利得をもつセミ・マルコフ決定過程について

長岡高専　涌田和芳（Kazuyoshi Wakuta）

§1. 序

非有界利得をもつセミマルコフ決定過程は最初 Lippman [6] により研究された。少し後で Lippman [7] は最初のモデルを一般化し，非可算状態空間であると同時により一般の推移法則をもつ場合を研究した。そこでは，weighted supremum norm を使って状態空間上の実数値関数の空間が Denardo [2] の N-stage contraction assumption を満足するような Banach space となるための十分条件を与えている。そして行動空間が有限のとき最適定常政策が存在することを示している。また，Nunen and Wessels [9] は Lippman の十分条件は少し弱いもので置きかえられることを示している。

ここでは，状態空間と行動空間がともに非可算集合である場合を考え，良く知られた selection theorem を適用して最適定常政策が存在するための十分条件について調べる。

§2. セミマルコフ決定過程

セミマルコフ決定過程は6つの組 (S, A, p, g, r, α) により定められる。S と A は Polish space の空でないボレル部分集合であり，それぞれ，状態空間と行動空間を表わす。$A(\cdot)$ は S から A への multifunction で，各 $s \in S$ に対して空でない実行可能な行動の集合を割合てる。p は K が与えられたときの S 上の regular conditional probability であり，システムの状態の運動法則を表わす。ここで，$K = \{(s,a) \mid a \in A(s)\}$ であり，K は $S \times A$ のボレル部分集合であると仮定する（この仮定については，後の注1で述べる）。g は $K \times S$ が与えられたときの R_+ 上の regular conditional probability であり，各状態での滞在時間の分布を表わす。ここで，$R_+ = [0, \infty)$ である。r は $K \times R \times S$ 上のボレル可測関数であり，利得を表わす。α は正数であり，割引因子を表わす。

もしシステムが決定を行なう時期に状態 $s \in S$ にあり，行動 $a \in A(s)$ が選択されるなら，(i) システムは分布 $p(\cdot \mid s, a)$ に従って選択される新しい状態 s' へ移る；(ii) 次の状態が s' であるという条件の下で，状態 s での滞在時間 τ は 分布 $g(\cdot \mid s, a, s')$ をもつ非負の確率変数である；(iii) そのとき，利得

$$r_\alpha(s,a,t,s') = \int_0^t r(s,a,t',s') e^{-\alpha t'} dt' \qquad (2.1)$$

を得る。

r の有界性のかわりに，p, q, r に次のような条件を課す。

Condition 1. (cf. Lippman [7], Nunen and Wessels [9])
任意の $s \in S$ と $a \in A(s)$ に対して，次の式を満たす S 上のボレル可測関数 $w(s) \geq 1$, ρ $(0 \leq \rho < 1)$, $M > 0$ が存在する：

$$\int_S \int_0^\infty |r_\alpha(s,a,t,s')| \, dq(t|s,a,s') \, dp(s'|s,a) \leq M w(s) \qquad (2.2)$$

and

$$\int_S \beta(s,a,s') w(s') \, dp(s'|s,a) \leq \rho w(s), \qquad (2.3)$$

ここで，

$$\beta(s,a,s') = \int_0^\infty e^{-\alpha t} dq(t|s,a,s') \qquad (2.4)$$

後のために

$$r(s,a) = \int_S \int_0^\infty r_\alpha(s,a,t,s') \, dq(t|s,a,s') \, dp(s'|s,a) \qquad (2.5)$$

とおく。ノルム

$$\|u\|_w = \sup_{s\in S} |u(s)|\, w(s)^{-1} \qquad (2.6)$$

をもつ S 上のボレル可測関数の Banach space を S^w と表わす。

初期状態 s が与えられて，政策 π が p，g とともに定める $A\times R_+\times S\times\cdots$ 上の条件付確率測度を $p_{\pi,s}$ と表わす。そして，任意の政策 π に対して，期待合計割引利得を

$$I(\pi)(s) = E_{\pi,s}\left[\sum_{n=1}^{\infty} e^{-\alpha T_{n-1}} r_\alpha(s_n, a_n, t_n, s_{n+1})\right] \qquad (2.7)$$

と定義する。ここで，$E_{\pi,s}[\cdot]$ は $p_{\pi,s}$ に関する条件付期待値を表わす。政策についての説明は省くが，すべての履歴に依存する確率的な政策を考えている。すべての政策 π とすべての状態 s に対して $I(\pi^*)(s) \geq I(\pi)(s)$ であるとき，政策 π^* を最適政策という。ここでの問題は，すべての政策の中で最適な定常政策が存在するための十分条件を調べることである。

次の propositions は，任意の政策 π と状態 s に対して，$p_{\pi,s}$ − almost surely に $\zeta \equiv \sup_{n\geq 1} T_n = \infty$ であり，I(π) は well-defined であることを言う。

Proposition 2.1. Condition 1 (2.3) を仮定する。このとき，任意の政策 π に対して，$p_{\pi,s}$ − almost surely に $\zeta \equiv \sup_{n\geq 1} T_n = \infty$ である。

(証明) 任意の $u \in S^w$ に対して,

$$E_{\pi,s}\left[e^{-\alpha T_{n-1}} |u(s_n)| \right] \leq \rho^{n-1} \|u\|_w w(s), \quad n \geq 1 \qquad (2.8)$$

が成立つことが帰納的に証明される。特に, $u \equiv 1$ とすると, (2.8)式より

$$E_{\pi,s}\left[e^{-\alpha T_{n-1}} \right] \leq \rho^{n-1} w(s) \longrightarrow 0 \quad (n \to \infty) \qquad (2.9)$$

したがって, $0 = T_0 \leq T_1 \leq \cdots$ なので, 有界収束定理より $\phi_{\pi,s}$ — almost surely に $\angle \neq \sup_{n \geq 1} T_n = \infty$ である。

Proposition 2.2. Condition 1 を仮定する。このとき, 任意の政策 π に対して, $\|I(\pi)\|_w \leq M/(1-\rho)$ である。

(証明) Proposition 2.1 と同様にして

$$E_{\pi,s}\left[e^{-\alpha T_{n-1}} |r_\alpha(s_n, a_n, t_n, s_{n+1})| \right] \leq \rho^{n-1} M w(s), \quad n \geq 1 \qquad (2.10)$$

を得る。したがって,

$$\begin{aligned} |I(\pi)(s)| &\leq E_{\pi,s}\left[\sum_{n=1}^{\infty} e^{-\alpha T_{n-1}} |r_\alpha(s_n, a_n, t_n, s_{n+1})| \right] \\ &\leq \left(M/(1-\rho) \right) w(s). \end{aligned}$$

ゆえに,

$$\| I(\pi) \|_w \leq M/(1-\beta).$$

Proposition 2.3. Condition 1 を仮定する。このとき, 任意の政策 π に対して,

$$I(\pi)(s) = E_{\pi,s}\left[\sum_{n=1}^{\infty} e^{-\alpha T_{n-1}} r(s_n, a_n) \right], \quad s \in S \qquad (2.11)$$

が成立つ。

(証明) 条件付期待値の性質より直ちに証明される。

§3. 最適定常政策

次の Assumptions をおく。

Assumption (a) (cf. Hinderer [5], Schäl [13]).

(i) S は Polish space の空でないボレル部分集合である;

(ii) A は locally compact separable metric space である;

(iii) $A(\cdot)$ は S から A への u.s.c. compact valued multifunction である;

(iv) 任意の continuous な $u \in S^w$ に対して

$$\int_S \beta(s,a,s') u(s') dp(s'|s,a)$$

は, K 上の continuous function である;

(v) rは, K上の u.s.c. functionである。

Assumption (b) (cf. Furukawa [3], Himmelberg et al. [4]).

(i) Sは Polish space の空でないボレル部分集合である;

(ii) Aは Polish space の空でないボレル部分集合である;

(iii) $A(\cdot)$はSからAへのボレル可測な compact valued multi-function である;

(iv) 任意の $u \in S^w$ に対して

$$\int_S \beta(s,a,s')\, u(s')\, d\phi(s'|s,a)$$

は, 各 $s \in S$ に対して $a \in A(s)$ の u.s.c. functionである;

(v) rは, 各 $s \in S$ に対して $a \in A(s)$ の u.s.c. fun. である。

(注1) Assumption (a), (b) a下で Kは $S \times A$ のボレル部分集合である。

最初に Assumption (a) の十分性について議論しよう。我々の問題でのいわゆる最適方程式は次のように表わされる (cf. Rosberg [10]):

$$u(s) = \max_{a \in A(s)} \left\{ r(s,a) + \int_S \beta(s,a,s')\, u(s')\, d\phi(s'|s,a) \right\}, \quad s \in S. \tag{3.1}$$

S^w のすべての u.s.c. functions のクラスを S_1^w と表わし, S_1^w 上に次のようなオペレーターを導入しよう:

$$T_a u(s) \equiv r(s,a) + \int_S \beta(s,a,s')\, u(s')\, dp(s'|s,a), \quad a \in A(s), \tag{3.2}$$

$$Tu(s) \equiv \max_{a \in A(s)} T_a u(s), \tag{3.3}$$

$$T_f u(s) \equiv r(s, f(s)) + \int_S \beta(s, f(s), s')\, u(s')\, dp(s'|s, f(s)) \tag{3.4}$$

ただし、f は $f(s) \in A(s)$, $s \in S$ なる S から A への ボレル 可測写像 である。

<u>lemma 3.1.</u> Condition 1 を仮定する。このとき、metric space $(S^w_1, \|\cdot\|_w)$ は complete である。ただし、w は u.s.c. と仮定する。

(証明) Maitra [8] と同様に 証明される。

<u>Condition 2.</u> $w \geq 1$ は S 上の continuous function である。

<u>lemma 3.2.</u> Condition 2 を仮定する。このとき、任意の $v \in S^w$ に対して、$v_n(s) \downarrow v(s)$, $s \in S$ なる continuous functions の列 $\{v_n\} \subset S^w$ が存在するときに限り、v は u.s.c. である。

(証明) 良く知られた u.s.c. functions に関する結果を少し修正して得られる (cf. Ash [1])。

lemma 3.3. Condition 1, 2, Assumption (a) を仮定する。このとき、任意の $v \in S_1^w$ に対して、

$$g(s,a) = \int_S \beta(s,a,s')\, v(s')\, dp(s'|s,a)$$

で定義される K 上の function g は、u.s.c. である。

(証明) lemma 3.2 から直ちに証明される。

lemma 3.4. Condition 1 と Assumption (a)(i), (ii), (iii) を仮定する。v は $|v(s,a)| \leq c\, w(s)$ (c は正の定数) なる K 上の u.s.c. function とする。このとき、

$$v^*(s) = \max_{a \in A(s)} v(s,a)$$

は、S_1^w に属する。

(証明) 明らかに $\|v^*\|_w \leq c$ である。v^* が u.s.c. であることは Schäl [13] と同様に証明される。

lemma 3.5. Condition 1, 2, Assumption (a) を仮定する。このとき、(3.3) で定義された オペレーター T は S_1^w 上の contraction mapping である。

(証明) 上の lemmas から直ちに証明される。

Theorem 3.6. Condition 1, 2, Assumption (a) を仮定する。このとき、最適方程式 (3.1) は一意の解 $u \in S_1^w$ をもち、

$$f(s) \in A(s) \tag{3.5}$$

and

$$u(s) = T_f u(s) = T u(s) \quad \text{for all } s \in S \tag{3.6}$$

なるボレル可測写像 $f : S \to A$ が存在する。

(証明) (3.5) と (3.6) を満たすボレル可測写像 f の存在を示せば良いが、これは Schäl [12] または Schäl [13] と同様に証明される。

Theorem 3.7. Condition 1 と Assumption (b) を仮定する。このとき、最適方程式 (3.1) は一意の解 $u \in S^w$ をもち、(3.5) と (3.6) を満たすボレル可測写像 $f : S \to A$ が存在する。

(証明) Himmelberg et al. [4] の結果を使って、上の lemma 3.1 — 3.5 と Theorem 3.6 と同様に証明される。

Theorem 3.8. Condition 1, 2, Assumption (a) (または Condition 1 と Assumption (b)) を仮定する。このとき、最適定常政策が存在し、最適利得は S_1^w (または S^w) に属する。

(証明) 過程の履歴を $h_n = (s, a_1, t_1, \cdots, s_{n-1}, a_{n-1}, t_{n-1}, s_n, a_n)$ とし，u は最適方程式 (3.1) の一意の解とする。このとき，任意の政策 π に対して，

$$E_{\pi,s}\left[\sum_{n=1}^{\infty}\left[e^{-\alpha T_n}u(s_{n+1}) - E_{\pi,s}\left[e^{-\alpha T_n}u(s_{n+1})\,|\,h_n\right]\right]\right] = 0 \quad (3.7)$$

が成立つ。ここで，$p_{\pi,s}$-almost surely に次のことが成立つ。

$$E_{\pi,s}\left[e^{-\alpha T_n}u(s_{n+1})\,|\,h_n\right]$$

$$= e^{-\alpha T_{n-1}}E_{\pi,s}\left[e^{-\alpha t_n}u(s_{n+1})\,|\,h_n\right]$$

$$= e^{-\alpha T_{n-1}}\int_S\int_0^\infty e^{-\alpha t}u(s')\,dq(t\,|\,s_n,a_n,s')\,d\phi(s'\,|\,s_n,a_n)$$

$$= e^{-\alpha T_{n-1}}\left[T_a u(s_n) - r(s_n,a_n)\right]$$

$$\leq e^{-\alpha T_{n-1}}\left[Tu(s_n) - r(s_n,a_n)\right] \quad (3.8)$$

$$= e^{-\alpha T_{n-1}}\left[u(s_n) - r(s_n,a_n)\right].$$

これを (3.7) 式へ代入すると，

$$E_{\pi,s}\left[e^{-\alpha T_N} u(s_{N+1}) - u(s) + \sum_{n=1}^{N} e^{-\alpha T_{n-1}} r(s_n, a_n) \right] \leqq 0 \qquad (3.9)$$

が成立つ。ここで、$N \to \infty$ とすると、(2.9) より

$$u(s) \geqq E_{\pi,s}\left[\sum_{n=1}^{\infty} e^{-\alpha T_{n-1}} r(s_n, a_n) \right] \qquad (3.10)$$

が成立つ。f を Theorem 3.6 または 3.7 で定義された ボレル可測写像とすると、定常政策 f^∞ は (3.8) 式、(3.10) 式で等号を満足する。したがって、

$$I(f^\infty)(s) = u(s) \geqq I(\pi)(s) \quad \text{for all } \pi \text{ and } s$$

が成立つ。ゆえに、定常政策 f^∞ は最適であり、u は最適利得である。

References

[1] Ash, R.B. (1972). Real Analysis and Probability. Academic Press.

[2] Denardo, E.V. (1967). Contraction mapping in the theory underlying dynamic programming. SIAM Review 9, 165-177.

[3] Furukawa, N. (1972). Markovian decision processes with compact action spaces. Ann. Math. Statist. 43, 1612–1622.

[4] Himmelberg, C.T., T. Parthasarathy, and F.S. Van Vleck (1976). Optimal plans for dynamic programming problems. Math. Oper. Res. 1. 390–394.

[5] Hinderer, K. (1970). Foundations of Non-Stationary Dynamic Programming with Discrete-Time Parameter. Springer-Verlag.

[6] Lippman, S.A. (1973). Semi-Markov decision processes with unbounded rewards. Management Sci. 19, 717–731.

[7] Lippman, S.A. (1975). On dynamic programming with unbounded rewards. Management Sci. 21, 1125–1233.

[8] Maitra, A. (1968). Discounted dynamic programming on compact metric spaces. Sankhya 30, Ser. A, 211–216.

[9] NunenVan, J. and J. Wessels (1978). A note on dynamic programming with unbounded rewards. Management Sci. 24, 576–580.

[10] Rosberg, Z. (1982). Semi-Markov decision processes with polynomial rewards. J. Appl. Prob. 19, 301–309.

[11] Ross, S.M. (1970). Average cost semi-Markov decision processes. J. Appl. Prob. 7, 649–659.

[12] Schäl, M. (1974). A selection theorem for optimization problem. Arch. Math. XXV, 219–224.

[13] Schäl, M. (1975). Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal. Z. Wahrscheinlichkeitstheorie Verw. Geb. 32, 179–196.